

DESEMPENHO DE MODELOS DE LINGUAGEM DE GRANDE ESCALA NO ACONSELHAMENTO PRÉ E PÓS-OPERATÓRIO PARA PROCEDIMENTOS FACIAIS ESTÉTICOS

João Marcos de Sales Brandão (joaomarcos0621@gmail.com)

Lucas de Araújo Albuquerque (lucasaraujoquerque@gmail.com)

Thais Muniz Adriano (thayzmuniz753@gmail.com)

Maria Eduarda Pereira Ponte (meduponteacademico@gmail.com)

Ariana Vasconcelos Aragão (arianavasconcelosaragao@gmail.com)

José Roberto Viana Silva (roberto_viana@yahoo.com)

Everton Freitas de Moraes (evertonfreitas2@gmail.com)

Introdução: Os modelos de linguagem de grande escala (LLMs) vêm sendo amplamente aplicados na área da saúde, porém seu uso em procedimentos cirúrgicos estéticos ainda é pouco explorado. Essas intervenções apresentam desafios singulares, marcados por elevadas expectativas dos pacientes, tomada de decisão emocionalmente sensível e resultados sutis, porém significativos, na autoimagem e no bem-estar psicossocial.

Objetivo: avaliar o desempenho de diferentes LLMs no aconselhamento pré e pós-operatório em cirurgia estética facial. Trata-se de um estudo transversal in silico que analisou o ChatGPT-4 (OpenAI, 2025), o DeepSeek V3 (DeepSeek AI/High-Flyer, 2025) e o Gemini 2.5 Pro Experimental (Google, 2025). **Metodologia:** Foram elaboradas vinte e seis perguntas padronizadas voltadas a pacientes, cujas respostas, anonimizadas, foram avaliadas de forma independente por dois cirurgiões bucomaxilofaciais calibrados. As análises contemplaram quatro domínios: precisão, empatia, legibilidade e confiabilidade das referências. Os dados foram submetidos aos testes Kruskal–Wallis, Mann–Whitney U com correção de Bonferroni, correlação de Spearman e qui-quadrado. **Resultados:** O modelo DeepSeek apresentou maior precisão ($4,77 \pm 0,51$; $p = 0,0078$) e legibilidade (FKRE $2,92 \pm 0,27$; $p < 0,00001$), enquanto o Gemini se destacou em empatia ($4,08 \pm 0,89$; $p < 0,001$). O GPT-4 produziu o maior número de citações alucinadas (36%), superando o Gemini (14%) e o DeepSeek (8,8%) ($p < 0,00001$). Foi observada correlação negativa entre empatia e legibilidade ($r = -0,34$; $p = 0,002$), indicando uma compensação



III SIMPÓSIO DE BIOTECNOLOGIA SEMINÁRIO

entre tom afetivo e clareza textual. **Conclusão:** Os LLMs analisados geraram respostas satisfatórias para o aconselhamento de pacientes, cada um com características específicas de desempenho, demonstrando potencial para aprimorar a comunicação centrada no paciente, embora exijam supervisão humana contínua para assegurar precisão e confiabilidade.

Palavras-chave: Medicina estética; Chatbots; Modelos de linguagem de grande escala; Inteligência Artificial